ITCS 6100 Big Data Analytics for Competitive Advantage
Group 1
Project Deliverable 1: Group Formation and Project Understanding

1. TEAM

TEAM MEMBERS

Delphine Antony Muthu - 801310101
Zahra Saghaie Dehkordi - 801212692
Kam Nanthanolath - 800983841

COMMUNICATION PLAN

- **Clarify roles and responsibilities:** Ensure that everyone in the team understands their specific role and responsibilities. This can help avoid confusion or duplication of work.
- **Method of Communication:** Utilize online tools and software for team communication, collaboration, and sharing of documents and files like Slack or Email
- **Regular team meetings:** Set up a schedule for regular team meetings to discuss progress, upcoming tasks, and any issues or concerns. This can be a weekly meeting where team members can share their updates and brainstorm ideas.
- **Clear deadlines:** Ensure that deadlines for tasks and projects are communicated clearly to all team members. This can help prevent delays and ensure that everyone is on the same page.
- **Monitor progress:** Regularly monitor progress on tasks and projects to ensure that they are on track. This can help us identify potential issues early on and allow for adjustments to be made before it's too late.
- **Meeting Attendance:** Meeting attendance is encouraged but not mandatory for all team members. If a team member is unable to attend, they should inform the team in advance.

PROJECT ARTIFACT REPOSITORY

Our project work can be found in the public repository that has been created on Github.
Link to the Repository - https://github.com/nthanol/6100Project

2. SELECTION OF DATASET (FIRST CHOICE)

From **Kaggle** dataset, *Heart Disease Prediction*:

https://www.kaggle.com/datasets/ritwikb3/heart-disease-cleveland

The dataset used for this task is the Cleveland Heart Disease dataset obtained from the UCI repository. The dataset comprises information from 303 individuals, and it includes 14 columns that were selected from a larger set of 75 columns. The objective of this classification task is to predict whether an individual has heart disease or not. The target variable has two possible values: 0, which indicates the absence of the disease, and 1, which indicates its presence.

3. BUSINESS PROBLEM, OPPORTUNITY AND DOMAIN KNOWLEDGE

BUSINESS PROBLEM:

The Cleveland Heart Disease dataset helps to assist healthcare providers in identifying patients who are at high risk of developing heart disease. By identifying individuals who are at risk of developing heart disease, healthcare providers can develop prevention strategies to reduce the risk of heart disease and related health complications. This model could also be used to prioritize patients for further diagnostic testing or treatments to prevent further problems.

DOMAIN KNOWLEDGE:

The Cleveland Heart Disease dataset lies in the field of cardiology and cardiovascular diseases. This dataset includes various attributes such as age, gender, blood pressure, cholesterol levels, presence of chest pain, etc. that are known risk factors for heart disease. Descriptive analytics can be performed on this dataset to identify the frequency of heart disease cases, and explore the different risk factors within the dataset. Predictive analytics can be applied to develop models that can predict whether a patient is likely to develop heart disease or not. This can help healthcare providers develop targeted prevention. Prescriptive analytics can be used to determine the best course of action to prevent heart disease in patients. This can involve developing personalized treatment plans based on a patient's risk factors and medical history, recommending lifestyle changes such as exercise and diet modifications, and determining the most effective medications to reduce a patient's risk of developing heart disease.

References:

https://archive.ics.uci.edu/ml/datasets/Heart+Disease

4. RESEARCH OBJECTIVES AND QUESTIONS

RESEARCH OBJECTIVE:

The research objective for the Cleveland Heart Disease dataset is to develop a predictive model that can accurately identify individuals who are at high risk of developing heart disease. The aim is to use the available information to create a model that can accurately predict the presence or absence of heart disease. This model can be used to identify individuals who require more aggressive treatment or lifestyle changes to reduce their risk of developing heart disease and prioritize patients for further testing and evaluation.

RESEARCH QUESTIONS:

The following research questions can be formulated to achieve the above objectives:

- What is the frequency distribution of heart disease present or not?
- How does the distribution of age vary between patients with and without heart disease?
- What is the relationship between the maximum heart rate achieved during exercise and heart disease?
- How does the distribution of fasting blood sugar vary between patients with and without heart disease?
- What is the relationship between chest pain type and heart disease?
- How does the distribution of thalassemia types vary between patients with and without heart disease?
- How does the distribution of cholesterol levels vary between patients with and without heart disease?
- What is the correlation between age and resting blood pressure?
- How do different risk factors for heart disease vary by age group?
- What is the distribution of age in the dataset?

2.SELECTION OF DATASET (SECOND CHOICE)

From **Kaggle** dataset, *FathomNet dataset - Image Classification of Marine Life*

https://www.kaggle.com/competitions/fathomnet-out-of-sample-detection

3.BUSINESS PROBLEM, OPPORTUNITY AND DOMAIN KNOWLEDGE

BUSINESS PROBLEM:

The business problem is that as researchers delve deeper into the ocean, their existing machine learning models begin to fail as new patterns in images appear. The depths are darker and unknown, which currently existing models weren't trained for. This can be addressed by utilizing this dataset to improve the efficiency of researchers' ability to identify new organisms. This overcomes the physical limits of the researchers and as an extension can improve efficiency as the researchers divert their efforts elsewhere.

DOMAIN KNOWLEDGE:

To address the issue of image classification, more technical skills in machine learning are required. Since the dataset consists of images, knowledge of convolutional neural networks and model architecture conventions is required to successfully complete the task.

- Convolutional Neural Networks - A type of neural network that trains using images. The model conducts convolutions on the image, detecting various features which allows better results on images than fully-connected networks.
- Residual Networks - Technique for neural networks that improves learning speed and accuracy potential. Distributes the machine learning input throughout the model which allows deeper layers to learn the input earlier. Used in popular architectures such as VGG19.

References:

- Suitable Loss Functions
- Domain Problem
- Data Source Explanation
- State of the Art Image Classification

4.RESEARCH OBJECTIVES AND QUESTIONS

RESEARCH OBJECTIVE:

- Create a model that generalizes well to new environments, organisms, and other image features
    - Create data transforms to corrupt the data to match deep-sea images
- Create a model that performs well in real-world applications
- Apply relevant state of the art techniques to the model
- Develop novel ideas to improve the model's performance

RESEARCH QUESTIONS:

- How many categories are there?
    - Is there enough data to represent each marine category?
- What architectural techniques are best for this situation?
- How do we properly transform the data to generalize well?
    - Would a GAN work?
- How much data is there?
- Can autoencoders be utilized to "denoise" the new image corruption?